We thank all reviewers for their valuable comments. Let us first provide a concise recap of our contributions. **i)** We derive a closed form expression of the minimizer of the squared risk under the Demographic Parity (DP) constraint (Thm. 2.3). **ii)** We propose an efficient post-processing algorithm (Alg. 1) which can be applied on top of *any* off-the-shelf estimator of the regression function, requiring *only* unlabeled data. **iii)** Our algorithm achieves strong finite sample fairness guarantees *without any* assumptions (Prop. 4.1). **iv)** Under additional assumptions, we derive plug-and-play finite sample risk guarantees (Thm. 4.4). These contributions lead to an intuitive understanding of DP (ll. 107–128), result in a computationally efficient method (ll. 156–162) which is interpretable and enjoys strong finite sample statistical guarantees (Section 4). We highlight that contributions **i)** and **iii)** are, up to our knowledge, unique.

We now address specific points raised by the reviewers, which will be included in the final version upon acceptance.

Optimal Transport (OT) + Fairness (R2, R4): Let us highlight two key differences between "Wasserstein Fair Classification" (Jiang et al.) and our work. **1.** While they directly work in the space of distributions and with transportation maps, we start from the problem of minimizing the risk under the DP constraint over *functions* and establish a link between the optimization over functions (*l.h.s.* problem in Thm. 2.3) and optimization over distributions (*r.h.s.* problem). In particular, they do not derive the form of a classifier which minimizes the *misclassification risk* (or the squared risk) under the DP constraint, a technical challenge that we solved in our paper for regression. **2.** Unlike our contribution, they neither provide risk guarantees nor they give bounds on the violation of the DP constraint, whereas we provide finite sample controls of both. Apart from shared spirit of OT, to the best of our knowledge, our contributions do not follow or generalize any previous work on fairness. On the other hand, our statistical analysis borrows tools from non-parametric statistics, rank statistics, empirical processes, and statistics in Wasserstein (Wass.) spaces.

Pareto frontier (R1, R3, R4): This is an interesting direction of future research. In order to study the Pareto frontier one needs to study the problem $\min\{\text{risk}(g) \,:\, \text{DP}(g) \leq \varepsilon\}$. Note that since DP is defined via the Kolmogorov-Smirnov (KS) distance it is oblivious to the geometry of the ambient space. In particular, the major technical challenge is to build a connection between the space of functions with $\mathbb{L}_2$ geometry and the distributions with the KS geometry. Our analysis establishes this connection for the case of DP $= 0$, by leveraging the fact that Wass. geometry in the space of distributions is "synchronized" with the squared risk geometry in the space of predictions. Another possible direction is to find g^*_ε, which minimizes the squared risk under the constraint that the Wass. barycenter objective is bounded by ε. Yet, this does not directly imply that g^*_ε is optimal for the problem $\min\{\text{risk}(g) \,:\, \text{DP}(g) \leq \varepsilon\}$.

Other notions of fairness (R3, R4): It would indeed be interesting to investigate extension of our analysis to other fairness notions. The main difficulty in such an extension for, e.g., Equalized Odds is due to the conditioning on the signal Y. Notice also that DP is used in several papers, including Jiang et al. discussed above.

R1. "naive" notion of fairness" Let us disagree that the notion of DP is naive. Generally group fairness constraints are trying to reflect a certain independence between the prediction and the sensitive attribute. DP is simply one of possible independence constraints that is, above all else, widely used in practice.

R2. "Assumption 4.2" As stated in the paper (ll. 196–202), we agree with R2 that As. 4.2 might be strong in certain situations. However a form of this assumption is rather classical in non-parametric statistics (see *e.g.,* "Fast learning rates for plug-in classifiers" Audibert & Tsybakov; Def. 2.2). In our settings As. 4.2 is mostly technical and can be further relaxed with much more involved analysis (see ll. 197–198). **"choice of sigma is left to the user."** We care to point out that Thm. 4.4 gives exact order of σ and Rem. 3.1 provides general guidelines. **"how to choose σ [...] why uniform noise?"** R2 raises an important point. Indeed, fairness guarantees (Prop. 4.1) do not require *any* condition on the noise level $\sigma{>}0$, while Thm. 4.4 gives its exact value. This discrepancy is dictated by completely different proof techniques of Prop. 4.1 and Thm. 4.4 and the fact that DP does not care about the quality of the base estimator. In particular, for Thm. 4.4 it is important that the noise: *i)* is continuous *ii)* does not deviate far from zero. Meanwhile, in Prop. 4.1 we only need the continuity of the noise and we do *not* care about its magnitude. Continuous noise allows us to derive *assumption free* fairness guarantees using tools from rank statistics and empirical processes. One can indeed use Gaussian noise with small variance. It does not affect Prop. 4.1 and the proof of Thm. 4.4 can be slightly modified.

R3. " [...] does not scale well to large number of sensitive features". We disagree with the reviewer. As indicated at ll. 160–161 our post-processing procedure has worst case *training* complexity $N \log N$ and $\log N$ for *inference* (with N being the total number of unlabeled data). **"[...] continuous sensitive attribute."** We thank the reviewer for this comment, it allowed us to extend our results to this case. Informally, it requires to replace $\mathbb{P}(S{=}s)$ by the density $\varphi(s)$ of random variable S ($\sum$ replaced by $\int$). Consequently, in the method (Eq. (6)) one needs to replace the estimates $\hat{p}_s$ by an estimator $\hat{\varphi}(s)$ of the density $\varphi(s)$ (*e.g.,* KDE). We will include this part in the final version. **"does this mean the probabilities are calculated in sample?"** Note that all of our bounds are *out of sample*. In Eq.(8) $\mathbf{P}$ stands for the joint distribution of data $\mathcal{D}$, added noise, and (X, S). Under $\mathbf{P}(\cdot|S{=}s, \mathcal{D})$, the method $\hat{g}$ is seen as non-random. Randomness comes only from the point (X, S). **"How do we know there aren't points [...] that Pareto dominate this method".** It is clear that the predictor g^* that *minimizes* the risk under the constraint that DP$=$0 is Pareto efficient, hence no other predictor can Pareto dominate g^*. Thanks to our finite sample guarantees, we can say that risk$(\hat{g}) \approx$ risk(g^*) and DP$(\hat{g}) \approx 0$. Thus $\hat{g}$ is nearly Pareto efficient and cannot be dominated by any other method at the population level.

R4. Given the above discussion, we hope that the reviewer is convinced that our contributions neither follow trivially from previous works on OT and fairness, nor can be seen as a straightforward extension to the regression setup.